ON BAYESIAN ESTIMATION AND PROXIMITY OPERATORS

R. GRIBONVAL AND M. NIKOLOVA

Abstract. There are two major routes to address the ubiquitous family of inverse problems appearing in signal and image processing, such as denoising or deblurring. A first route relies on Bayesian modeling, where prior probabilities are used to embody models of both the distribution of the unknown variables and their statistical dependence with respect to the observed data. The estimation process typically relies on the minimization of an expected loss (e.g. minimum mean squared error, or MMSE). The second route has received much attention in the context of sparse regularization and compressive sensing: it consists in designing (often convex) optimization problems involving the sum of a data fidelity term and a penalty term promoting certain types of unknowns (e.g., sparsity, promoted through an ℓ^1 norm).

Well known relations between these two approaches have led to some widely spread misconceptions. In particular, while the so-called Maximum A Posterori (MAP) estimate with a Gaussian noise model does lead to an optimization problem with a quadratic data-fidelity term, we disprove through explicit examples the common belief that the converse would be true.

It has already been shown [7, 9] that for denoising in the presence of additive Gaussian noise, for *any* prior probability on the unknowns, MMSE estimation can be expressed as a penalized least squares problem, with the apparent characteristics of a MAP estimation problem with Gaussian noise and a (generally) different prior on the unknowns. In other words, the variational approach is rich enough to build all possible MMSE estimators associated to additive Gaussian noise via a well chosen penalty.

We generalize these results beyond Gaussian denoising and characterize noise models for which the same phenomenon occurs. In particular, we prove that with (a variant of) *Poisson* noise and any prior probability on the unknowns, MMSE estimation can again be expressed as the solution of a penalized least squares optimization problem. For *additive* scalar denoising the phenomenon holds if and only if the noise distribution is log-concave. In particular, Laplacian denoising can (perhaps surprisingly) be expressed as the solution of a penalized least squares problem. In the multivariate case, the same phenomenon occurs when the noise model belongs to a particular subset of the exponential family. For multivariate *additive* denoising, the phenomenon holds if and only if the noise is white and Gaussian.

This work and the companion paper [10] are dedicated to the memory of Mila Nikolova, who passed away prematurely in June 2018. Mila dedicated much of her energy to bring the technical content to completion during the spring of 2018. The first author did his best to finalize the papers as Mila would have wished. He should be held responsible for any possible imperfection in the final manuscript.

R. Gribonval, Univ Rennes, Inria, CNRS, IRISA, remi.gribonval@inria.fr;
M. Nikolova, CMLA, CNRS and Ecole Normale Supérieure de Cachan, Université Paris-Saclay, 94235 Cachan, France.

1. Introduction and overview

Inverse problems in signal and image processing consist in estimating an unknown signal x_0 given an indirect observation y that may have suffered from blurring, noise, saturation, etc. The two main routes to address such problems are variational approaches and Bayesian estimation.

Variational approaches: a signal estimate is the solution of an optimization problem

$$\hat{x} \in \arg\min_x D(x,y) + \varphi(x) \tag{1}$$

where $D(x,y)$ is a data-fidelity measure, and φ is a penalty promoting desirable properties of the estimate $\hat{x}$ such as, e.g., sparsity.

A typical example is linear inverse problems, where one assumes $y = Lx_0 + e$ with L some known linear operator (e.g., a blurring operator), e is some error / noise. The most common data-fidelity term is the square of the Euclidean norm $\|\cdot\|$, which in combination with an ℓ^1 sparsity-enforcing penalty leads to the well known Basis Pursuit Denoising approach

$$\hat{x}_{\mathrm{BPDN}}(y) := \arg\min_x \tfrac{1}{2}\|y - Lx\|^2 + \lambda\|x\|_1. \tag{2}$$

Bayesian estimation: x_0 is modeled as a realization of a random variable X (with "prior" probability $p_X(x)$) and y as the realization of a random variable Y (with conditional probability distribution $p_{Y|X}(y|x)$). A Bayesian estimator is designed as a function $y \mapsto \hat{x}(y)$ that minimizes in expectation some specified cost $C(x_0, \hat{x})$, i.e., that minimizes

$$\mathbb{E}_{X,Y}\ C(X, \hat{x}(Y)) \tag{3}$$

where the pair (X,Y) is drawn according to the joint distribution $p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$. Equivalently, for a given y, the estimator $\hat{x}(y)$ is a minimizer of $\mathbb{E}_{X|Y=y}C(X, \hat{x}(y))$.

A typical example is Minimum Mean Square Error (MMSE) estimation. The cost is the quadratic error $C(x,\hat{x}) = \|\hat{x} - x\|^2$, and the optimal estimator is the conditional expectation, also called Conditional Mean or Posterior Mean

$$y \mapsto \hat{x}_{\mathrm{MMSE}}(y) := \mathbb{E}(X|Y=y) = \int x p_{X|Y}(x|Y=y)dx. \tag{4}$$

By Bayes law $p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{p_Y(y)}$, with $p_Y(y)$ the marginal distribution of Y.

While MMSE estimation yields the *expected value* of the *a posteriori* probability distribution $p_{X|Y}(x|y)$ of X, Maximum A Posteriori (MAP) estimation selects its mode, i.e. the *most probable* x with respect to this distribution,

$$\hat{x}_{\mathrm{MAP}}(y) := \arg\min_x -\log p_{X|Y}(x|y) = \arg\min_x \{-\log p_{Y|X}(y|x) - \log p_X(x)\}.$$

MAP is directly connected to variational approaches, hence its popularity. However, this is not usually considered as a proper Bayesian estimator although it can be seen as minimizing (3) with the "pseudo-cost" $C(x,\hat{x}) := \delta(x - \hat{x})$. We will soon come back to the Bayesian interpretation of MAP estimation and its pitfalls.

Many other costs can be used, e.g. $C(x,\hat{x}) = \|x - \hat{x}\|$ yields the conditional spatial median.

Banerjee et al [3] show that if $C(x,\hat{x})$ (defined on $\mathbb{R}^n \times \mathbb{R}^n$) is the Bregman divergence[1] $D_h(x,\hat{x})$ of a strictly convex proper differentiable function h [5], then the conditional mean is the unique minimizer of (3) [3, Theorem 1]. Vice-versa, they prove under mild smoothness assumptions on $C(x,\hat{x})$ that if the conditional mean is the unique minimizer of (3) for any pair of random variables X, Y then

$$C(x,\hat{x}) = D_h(x,\hat{x}), \quad \forall\, x,\hat{x} \tag{5}$$

for some strictly convex differentiable function h.

1.1. **The MAP vs MMSE quid pro quo.** A common quid pro quo between tenants of the two approaches revolves around the MAP interpretation of variational approaches [12]. In the particular case of a linear inverse problem with white Gaussian noise the conditional density reads[2] $p_{Y|X}(y|x) \propto \exp\left(-\frac{\|y-Lx\|^2}{2\sigma^2}\right)$, and denoting $\varphi_X(x) := -\sigma^2 \log p_X(x)$, MAP estimation reads

$$\hat{x}_{\mathrm{MAP}} = \arg\min_x \tfrac{1}{2\sigma^2}\|y-Lx\|^2 - \log p_X(x) = \arg\min_x \tfrac{1}{2}\|y-Lx\|^2 + \varphi_X(x) \tag{6}$$

Thus, *if* one assumes a Gaussian noise model and *if* one chooses MAP as an estimation principle *then* this results in a variational problem shaped as (1) with a quadratic data-fidelity term and a penalty which is a scaled version of the negative log-prior.

It is argued in [17] that –except in very special circumstances (Gaussian prior and Gaussian noise)– "*the MAP approach is not relevant in the applications where the data-observation and the prior models are accurate*", in that (6) leads to a suboptimal estimator when the considered data is indeed distributed as $p_X(x) \propto \exp(-\varphi_X(x)/\sigma^2)$ with Gaussian noise. Indeed, consider as an example compressive sensing where $y = Lx + e$ with Gaussian i.i.d. e and L an underdetermined measurement matrix. When X is a random vector with i.i.d. Laplacian entries, we get $\varphi_X(x) \propto \|x\|_1$ and the MAP estimator is Basis Pursuit Denoising (2) which has been shown [8] to have poorer performance (in the highly compressed regime and in the limit of low noise) than a variational estimator (1) with quadratic data-fidelity and quadratic penalty $\varphi(x) \propto \|x\|_2^2$, aka Tikhonov regression or ridge regression.

Unfortunately, a widely spread misconception has led to a "reverse reading" of optimization problems associated to variational approaches. For example, even though it is true that one obtains (2) as a MAP under additive Gaussian noise with a Laplacian signal prior, by no means does this imply that the use of (2) to build an estimator is necessarily motivated by the *choice* of MAP as an estimation principle and the *belief* that the Laplacian prior is a good description of the distribution of X. Instead, as widely documented in the literature on sparse regularization, the main reason for choosing the ℓ^1 penalty is simply to promote sparse solutions: any minimizer of (2) is bound to have (many) zero entries (in particular when the parameter λ is large) which is desirable when prior knowledge indicates that x_0 is "compressible", that is to say, well approximated by a sparse vector.

As demonstrated e.g. in [8] (see also [2] for related results), a random vector $X \in \mathbb{R}^n$ with entries drawn i.i.d. from a Laplacian distribution is, with high probability, *not* compressible; on

[1] By definition $D_h(x,y) := h(x) - h(y) - \langle \nabla h(y), x-y\rangle$. This is usually not symmetric, i.e. $D_h(x,y) \neq D_h(y,x)$.

[2] with $\propto$ denoting proportionality.

the opposite, when the entries of X are drawn i.i.d. according to a *heavy-tailed distribution*, X is typically compressible. Thus, despite having the apparent characteristics of the MAP estimator with a Laplacian prior, (2) in fact approximates well the MMSE estimator with a heavy-tailed prior.

REMARK 1. *One should be careful about the temptation of determining once and for all which of MAP or MMSE is a "better" estimation principle. While the above compressive sensing example illustrates that MAP estimation under the true underlying distribution can lead to poorer estimation performance than MMSE, J. Idier (private correspondence) provided the authors with an interesting example of the converse phenomenon, where MMSE estimation is not as operational as MAP. This happens, e.g. in phase unwrapping where x needs to be estimated modulo 2π. MAP then involves finding one of the modes of the posterior $p_{X|Y}(x|y)$, all modes being more or less equal modulo 2π, hence equally operational. On the contrary MMSE typically averages several modes and leads to non-operational estimates. Similar problems may arise in other estimation problems where label switching may occur, leading to multiple modes of the posterior distribution.*

REMARK 2. *In* [17]*, an argument against the MAP approach is expressed as:* "In full rigor, an estimator $\hat{X}$ for X, based on data Y, can be said to be coherent with the underlying models if $\hat{X} \sim f_X$"*, i.e., if $\hat{X}$ has the same probability distribution as X. Again, as pointed out by J. Idier, such a criterion to qualify a "good" estimator seems questionable: while "$\hat{X} \sim f_X$" can be (approximately) expected when the knowledge of Y allows to perfectly estimate X (or, on the opposite, when Y brings no information on X!), such a property cannot be expected in the more realistic intermediate cases: even in the linear Gaussian setting, where MAP and MMSE lead to the same (unbiased) estimator, the covariance of the estimator generally does not match that of the prior. A more convincing intrinsic issue with MAP is probably its lack of invariance with respect to reparametrization of the problem.*

1.2. **Writing certain *convex* variational estimators as proper Bayes estimators.** In penalized least squares regression for linear inverse problems in $\mathbb{R}^n$, the variational estimator

$$\hat{x}(y) \in \arg\min_{x\in\mathbb{R}^n} \tfrac{1}{2}\|y - Lx\|^2 + \varphi(x), \tag{7}$$

is traditionally interpreted as a MAP estimator (6) with Gaussian noise and prior

$$p_X(x) \propto \exp(-\varphi(x)).$$

When the penalty φ is convex differentiable and $\frac{1}{2}\|Lx\|^2 + \varphi(x)$ has at least linear growth at infinity, by [6, Theorem 1] the estimator (7) is also a proper Bayesian estimator minimizing (3), with a prior distribution $p_X(x) \propto \exp(-\varphi(x))$ and $y = Lx + e$ where e is white Gaussian noise, for the cost

$$C(x, \hat{x}) = \tfrac{1}{2}\|L(x - \hat{x})\|^2 + D_\varphi(\hat{x}, x) \tag{8}$$

involving the Bregman divergence D_φ. The results in [6] are actually expressed with colored Gaussian noise and with $\|\cdot\|$ replaced by the corresponding weighted (Mahalanobis) quadratic norm in (7) and (8). Further extensions to infinite-dimension have been considered in [11].

In other words, in the context of linear inverse problems with additive Gaussian noise and a log-concave prior, a variational approach (7) having all the apparent features of the MAP estimator is in fact also a proper Bayesian estimator with the specific cost (8).

REMARK 3. The reader may notice that the cost (8) is a Bregman divergence $C(x,\hat{x}) = D_h(\hat{x},x)$, with $h(x) := \frac{1}{2}\|Lx\|^2+\varphi(x)$. The order of arguments is reversed compared to (5), which is neither an error nor a coincidence: by the results of Banerjee [3], when $C(x,\hat{x}) = D_h(x,\hat{x})$ for some h, the Bayes estimator minimizing (3) is simply the MMSE. Here, in (7)-(8), this does not happen: the MAP estimator does not match the MMSE estimator except in certain very specific cases where the Bregman divergence is indeed symmetric.

1.3. **Writing certain MMSE estimators using an expression analog to a MAP estimator.** The results of Burger et al strongly intertwine the prior model $p_X(x) \propto e^{-\varphi(x)}$, the observation model $p_{Y|X}(y|x)$ embodied by the linear operator L, and the task cost $C_{L,\varphi}(x,\hat{x})$. In particular, the task ("what we want to do") becomes *dependent* on the data model ("what we believe").

From a data processing perspective the above approach is not fully satisfactory: it seems more natural to first choose a relevant task (e.g., MMSE estimation) and a reasonable model (e.g., a prior $p_X(x)$), and then to design a penalty (the tool used to solve the task given the model) based on these choices.

In this spirit, in the context of additive white Gaussian denoising, [7] showed that, *for any signal prior* $p_X(x)$, the MMSE estimator (4) is the unique solution (and unique stationary point) of a variational optimization problem

$$\hat{x}_{\mathrm{MMSE}}(y) = \arg\min_x \tfrac{1}{2}\|y-x\|^2 + \widetilde{\varphi}_X(x) \tag{9}$$

with $\widetilde{\varphi}_X(x)$ some penalty that depends on the considered signal prior $p_X(x)$.

In other words, MMSE has all the apparent features of a MAP estimator with Gaussian noise and a "pseudo" signal prior $\widetilde{p}_X(x) \propto e^{-\widetilde{\varphi}_X(x)}$. Except in the very special case of a Gaussian prior, *the pseudo-prior* $\widetilde{p}_X(x)$ *differs from the prior* $p_X(x)$ that defines the MMSE estimator (4). This result has been extended to MMSE estimation for inverse problems with additive colored Gaussian noise [9]. Unser and co-authors [13, 1],[18, Section 10.4.3] exploit these results for certain MMSE estimation problems. Louchet and Moisan [14, Theorem 3.9] consider the specific case of the MMSE estimator associated to a total variation image prior $p_X(x)$ and establish the same property through connections with the notion of *proximity operator* of a convex lsc function.

1.4. **Contribution: MMSE estimators that can be expressed as proximity operators.** We extend the general results of [8, 9] beyond Gaussian denoising using a characterization of proximity operators of possibly nonconvex penalties obtained in a companion paper [10]. Our extension goes substantially beyond Gaussian denoising, including scalar Poisson denoising, scalar denoising in the presence of additive noise with any log-concave distribution, and multivariate denoising for certain noise distributions belonging to the exponential family.

1.4.1. *Scalar denoising.*

PROPOSITION 1 (scalar Poisson denoising). *Consider the scalar Poisson noise model where the conditional probability distribution of the integer random variable $Y \in \mathbb{N}$ given $x \in \mathbb{R}_+^*$ is*

$$p_{Y|X}(Y = n|x) = \tfrac{x^n}{n!} e^{-x}, \quad \forall\, n \in \mathbb{N}$$

and let p_X be any probability distribution for a positive random variable $X > 0$. There is a (possibly nonconvex) penalty function $\tilde{\varphi}_X : \mathbb{R}_+^* \to \mathbb{R} \cup \{+\infty\}$ such that, for any $n \in \mathbb{N}$

$$\mathbb{E}(X|Y = n) \;\in\; \arg\min_{x>0} \left\{ \tfrac{1}{2}(n - x)^2 + \tilde{\varphi}_X(x) \right\}. \tag{10}$$

Note that the arg min *is strictly positive since the conditional expectation is strictly positive. The penalty $\tilde{\varphi}_X$ depends on the probability distribution p_X of X.*

PROPOSITION 2 (scalar additive noise). *Consider an* additive *noise model $Y = X + N$ where the random variables $X, Z \in \mathbb{R}$ are independent. The conditional probability distribution of the random variable $Y \in \mathbb{R}$ given $x \in \mathbb{R}$ is $p_{Y|X}(y|x) = p_Z(y - x)$. Assume that $p_Z(z) > 0$ for any $z \in \mathbb{R}$ and that $z \mapsto F(z) := -\log p_Z(z)$ is continuous. The following properties are equivalent:*

(a) *the function F is convex (i.e., the noise distribution is log-concave);*
(b) *for any prior probability distribution p_X on the random variable $X \in \mathbb{R}$, the conditional expectation $\mathbb{E}(X|Y = y)$ is well defined for any $y \in \mathbb{R}$, and there is a (possibly nonconvex) penalty function $\tilde{\varphi}_X : \mathbb{R}_+^* \to \mathbb{R} \cup \{+\infty\}$ such that for any $y \in \mathbb{R}$*

$$\mathbb{E}(X|Y = y) \;\in\; \arg\min_{x\in\mathbb{R}} \left\{ \tfrac{1}{2}(y - x)^2 + \tilde{\varphi}_X(x) \right\}. \tag{11}$$

The penalty $\tilde{\varphi}_X$ depends on the probability distribution p_X of X and the function F.

Proposition 1 and Proposition 2 are proved in Section 2 as corollaries of a more general result (Lemma 1) on scalar MMSE estimation. Let us insist that while the naive interpretation of an optimization problem such as (10) (resp. (11)) would be that of MAP estimation with Gaussian noise, it actually corresponds to MMSE estimation with Poisson (resp. log-concave) noise.

Classical log-concave examples include generalized Gaussian noise [15, 4], where $p_{Y|X}(y|x) \propto \exp\left(-\left|\frac{x-y}{\sigma}\right|^\gamma\right)$ with $1 \leqslant \gamma < \infty$. This includes Gaussian noise for $\gamma = 2$, but also Laplacian noise for $\gamma = 1$, and in all cases the MMSE estimator can always be written as a "pseudo-MAP" with a "Gaussian" (i.e., quadratic) data-fidelity term and an adapted "pseudo-prior" $\tilde{\varphi}_X$. For $0 < \gamma < 1$ the noise distribution is not log-concave, hence *there are prior probability distributions p_X such that the corresponding MMSE estimator* cannot be written as in (11).

EXAMPLE 1. Consider X a scalar random variable with a Laplacian prior distribution, and $Y = X + cZ$ where Z is an independent scalar Laplacian noise variable and $c > 0$. While the MAP estimator reads

$$\arg\min_x \{|y - x| + c|x|\},$$

by Proposition 2 the MMSE estimator can be expressed as

$$f(y) = \arg\min_x \left\{ \tfrac{1}{2}(y - x)^2 + \varphi(x) \right\}$$

with some penalty φ. The details of the analytic derivation of f are in Appendix A.8. The corresponding potential ψ (cf Theorem 1 in Section 2) can be obtained by numerical integration,

and the penalty φ is characterized by Eq. (16): $\varphi(f(y)) = yf(y) - f^2(y)/2 - \psi(y)$. It can be plotted using the pairs $f(y)$,$\varphi(f(y))$. Figure 1 provides the shape of $f(y)$, of the corresponding potential $\psi(y)$, and of the corresponding penalty $\varphi(x)$ for $c = 0.9$.

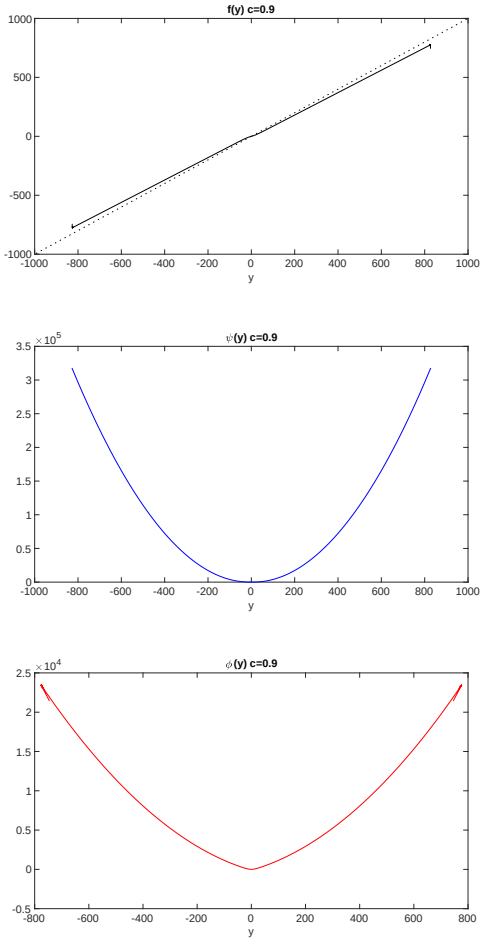

FIGURE 1. $f(y)$ (top), potential $\psi(y)$ from Theorem 1 (middle) and penalty $\varphi(y)$ (bottom) for Example 1 with $c = 0.9$.

1.4.2. *Multivariate denoising.* The above scalar examples straightforwardly extend to the multivariate setting in the special case where both the noise model and the prior p_X are completely

separable, cf Example 2 in Section 2. In general however, the multivariate setting is quite different from the scalar one. For example in $\mathbb{R}^n$, $n \geqslant 2$, MMSE estimation in the presence of additive Laplacian noise (resp. of Poisson noise) *cannot always* be written as in (9), depending on the prior p_X, see Example 3 (resp. Example 4) in Section 2. In contrast, we show (Lemma 5) that the MMSE estimator can always be expressed as in (9), provided we consider particular noise models of the exponential family

$$p_{Y|X}(y|x) = \exp\left(c\langle x, y\rangle - a(x) - b(y)\right) \tag{12}$$

for some $c \geqslant 0$ and smooth enough b. This form is in fact essentially necessary (Lemma 4).

A primary example is additive Gaussian white noise, where

$$p_{Y|X}(y|x) = C\exp\left(-\tfrac{\|y-x\|^2}{2\sigma^2}\right) = \exp\left(\log C - \tfrac{\|y\|^2}{2\sigma^2} - \tfrac{\|x\|^2}{2\sigma^2} + \tfrac{1}{\sigma^2}\langle x, y\rangle\right), \tag{13}$$

and we recover the main results of [7], using the following definition.

Definition 1. *A random variable X with values in a Hilbert space $\mathcal{H}$ and probability distribution P is non-degenerate if there is no affine hyperplane $\mathcal{V} \subset \mathcal{H}$ such that $X \in \mathcal{V}$ almost surely, i.e., if $P(\langle X, v\rangle = d) < 1$ for all nonzero $v \in \mathcal{H}$ and $d \in \mathbb{R}$.*

Proposition 3 (additive white Gaussian denoising). *Consider an additive noise model $Y = X + Z \in \mathbb{R}^n$ where the random variables X, Z are independent and $p_Z(z) \propto \exp\left(-\frac{\|z\|^2}{2\sigma^2}\right)$. Then for any prior distribution p_X there is a (possibly nonconvex) penalty $\widetilde{\varphi}_X : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ such that for any $y \in \mathbb{R}^n$*

$$\mathbb{E}(X|Y = y) \;\in\; \arg\min_{x\in\mathbb{R}^n} \{\tfrac{1}{2}\|y - x\|^2 + \widetilde{\varphi}_X(x)\}. \tag{14}$$

The penalty $\widetilde{\varphi}_X$ depends on the probability distribution p_X of X and the variance σ^2.

Further, if X is non-degenerate then the function $y \mapsto \mathbb{E}(X|Y = y)$ is injective and there is a choice of $\widetilde{\varphi}_X$ such that for any y, $\mathbb{E}(X|Y = y)$ is the unique stationary point (and global minimizer) of the rhs of (14), hence $y \mapsto \mathbb{E}(X|Y = y)$ is the proximity operator of $\widetilde{\varphi}_X$.

Remark 4. *It is further known [7] that in this Gaussian context the conditional mean estimator $y \mapsto \mathbb{E}(X|Y = y)$ is C^∞, and that $e^{-\widetilde{\varphi}_X(x)}$ is (up to renormalization) a proper prior density. It is also known [9] that $\widetilde{\varphi}_X$ is convex (resp. additively separable) if and only if the marginal $p_Y(y)$ is log-concave (resp. separable); this holds in particular as soon as the prior p_X is log-concave (resp. separable). Extending such characterizations beyond Gaussian denoising is postponed to further work.*

This is the only example of centered additive noise with smooth density of the form (12).

Proposition 4. *Consider a multivariate* centered additive *noise model i.e. with $p_{Y|X}(y|x) = p_Z(y - x)$. Assume that $p_Z(z) > 0$ for any $z \in \mathcal{H}$ and denote $z \mapsto F(z) := -\log p_Z(z)$. If F is continuous and $p_{Y|X}$ is of the form (12) then $p_{Y|X}(y|x)$ is in fact Gaussian, with the expression (13) for some $\sigma > 0$.*

The proof is in Appendix A.6. Despite Proposition 4 and the apparent connection between the quadratic fidelity term in (14) and the Gaussian log-likelihood $-\log p_{Y|X}(y|x)$, noise models

of the form (12) are more general. For example, they cover a variant of multivariate Poisson denoising.

PROPOSITION 5 (variant of multivariate Poisson denoising). *Consider a random variable $Y \in \mathbb{N}^n$ with conditional distribution given $x \in (\mathbb{R}_+^*)^n$ expressed as*

$$p_{Y|X}(Y = y|x) = \prod_{i=1}^{n} \frac{x_i^{y_i}}{y_i!} e^{-x_i}, \quad \forall\, y \in \mathbb{N}^n,$$

and let p_X be any probability distribution for a multivariate positive random variable X. There is a (possibly nonconvex) penalty function $\tilde{\varphi}_X : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ such that the MMSE estimator of the (entrywise) logarithm $\log X = (\log X_i)_{i=1}^n$ *satisfies for any $y \in \mathbb{N}^n$*

$$\mathbb{E}(\log X|Y = y) \;\in\; \arg\min_{\xi \in \mathbb{R}^n} \left\{ \tfrac{1}{2}\|y - \xi\|^2 + \tilde{\varphi}_X(\xi) \right\}. \tag{15}$$

The penalty $\tilde{\varphi}_X$ depends on the probability distribution p_X of X.

Further, if X is non-degenerate then the function $y \mapsto \mathbb{E}(X|Y = y)$ is injective and there is a choice of $\widetilde{\varphi}_X$ such that for any y, $\mathbb{E}(X|Y = y)$ is the unique stationary point (and global minimizer) of the rhs of (15)*, hence $y \mapsto \mathbb{E}(X|Y = y)$* is *the proximity operator of $\widetilde{\varphi}_X$.*

In Proposition 5 the noise model is not Gaussian, yet MMSE estimation can be expressed in a variational form "looking like" Gaussian denoising due to the quadratic data-fidelity term in (15). Proposition 3 and Proposition 5 are proved in Section 2 as corollaries of Lemma 5.

1.5. **Discussion and extensions.** While this paper primarily focuses on characterizing when the conditional mean estimator is a proximity operator, it is natural to wonder under which noise models the conditional mean estimator can be expressed as the solution of another variational optimization problem such as (2) or more generally (1), with properly chosen data fidelity term and penalty. This has been done [9] for linear inverse problems with colored Gaussian noise, and we expect that it is possible to combine the corresponding proof techniques with [10, Theorem 3(c) and Corollary 6] to obtain results in this vein for a larger class of observation models. Of particular interest would be to understand whether Poisson denoising can be written as the solution of a variational problem (1) with a Kullback-Leibler divergence as data-fidelity term (which appears naturally in a MAP framework) and a well chosen penalty.

Finally, the reader may have noticed that the characterizations obtained in this paper are non constructive. They merely state the *existence* of a penalty such that the MMSE, which is a priori expressed as an integral, is in fact the solution of a (often convex) variational problem. From a practical point of view, a challenging perspective would be to identify how to exploit this property to design efficient estimation algorithms.

Acknowledgement. The corresponding author is indebted to the anonymous reviewers for their remarks, as well as to J. Idier for his insightful comments on a preliminary version of this work which have led in particular to Remarks 1 and 2.

2. When is MMSE estimation a proximity operator ?

We now provide our main general results on the connections between Bayesian estimation and variational approaches. After some reminders on the expression of the conditional mean

in terms of a marginal and a conditional distribution, we define a class of "proto" conditional means estimators based on a given "proto" conditional distribution $q(y\,|\,x)$. Then, we focus on the scalar case where we characterize (proto) conditional distributions that lead to (proto) conditional mean estimators that are proximity operators. Finally we consider the vector case.

2.1. **Reminders on proximity operators.** [11]

Let $\mathcal{H}$ be a Hilbert space equipped with an inner product denoted $\langle\cdot,\cdot\rangle$ and a norm denoted $\|\cdot\|$. In the finite dimensional case, one may simply consider $\mathcal{H}=\mathbb{R}^n$. A function $\theta:\mathcal{H}\to\mathbb{R}\cup\{+\infty\}$ is proper iff there is $x\in\mathcal{H}$ such that $\theta(x)<+\infty$, i.e., $\mathrm{dom}(\theta)\neq\varnothing$, where $\mathrm{dom}(\theta):=\{x\in\mathcal{H}\mid\theta(x)<\infty\}$. It is lower semi-continuous (lsc) if for any $x_0\in\mathcal{H}$, $\liminf_{x\to x_0}\theta(x)\geqslant\theta(x_0)$, or equivalently if the set $\{x\in\mathcal{H}:\theta(x)>\alpha\}$ is open for every $\alpha\in\mathbb{R}$. The proximity operator of a (possibly nonconvex) proper penalty function $\varphi:\mathcal{H}\to\mathbb{R}\cup\{+\infty\}$ is the set-valued operator

$$y\mapsto\mathrm{prox}_{\varphi}(y):=\arg\min_{x\in\mathcal{H}}\left\{\tfrac{1}{2}\|y-x\|^2+\varphi(x)\right\}$$

A primary example is soft-thresholding $f(y):=y(1-1/y)_+$, $y\in\mathcal{H}:=\mathbb{R}$, which is the proximity operator of the absolute value function $\varphi(x):=|x|$. Let us recall that if θ is a convex function, $u\in\mathcal{H}$ is a subgradient of θ at $x\in\mathcal{H}$ iff $\theta(x')-\theta(x)\geqslant\langle u,x'-x\rangle$, $\forall\,x'\in\mathcal{H}$. The subdifferential at x is the collection of all subgradients of θ at x and is denoted $\partial\theta(x)$.

The following results are proved in the companion paper [10]. Strictly speaking, a proximity operator is set-valued: a function f such that $f(y)\in\mathrm{prox}_{\varphi}(y)$ for any $y\in\mathcal{Y}$ is a *selection* of the proximity operator of φ. For concision we will say that f *implements* this proximity operator.

Theorem 1. [10, Theorem 1] *Consider $\mathcal{Y}$ a non-empty subset of $\mathcal{H}$. A function $f:\mathcal{Y}\to\mathcal{H}$ implements the proximity operator of some penalty φ (i.e. $f(y)\in\mathrm{prox}_{\varphi}(y)$ for any $y\in\mathcal{Y}$) iff there exists a convex lsc function $\psi:\mathcal{H}\to\mathbb{R}\cup\{+\infty\}$ such that for any $y\in\mathcal{Y}$, $f(y)\in\partial\psi(y)$.*

When the domain $\mathcal{Y}$ is convex, [10, Theorem 3] implies that there is a number $K\in\mathbb{R}$ such that the functions f, φ and ψ in Theorem 1 satisfy

$$\psi(y)=\langle y,f(y)\rangle-\tfrac{1}{2}\|f(y)\|^2-\varphi(f(y))+K,\quad\forall\,y\in\mathcal{Y}. \tag{16}$$

Corollary 1. [10, Corollary 4] *Consider an arbitrary non-empty subset $\mathcal{Y}\subset\mathbb{R}$. A function $f:\mathcal{Y}\to\mathbb{R}$ implements the proximity operator of some penalty φ if, and only if, it is nondecreasing.*

Theorem 2. [10, Theorem 2] *Let $\mathcal{Y}$ be an open convex subset of $\mathcal{H}$ and $f:\mathcal{Y}\to\mathcal{H}$ be $C^1(\mathcal{Y})$. The following properties are equivalent:*

(a) *f implements the proximity operator of a function φ (i.e. $f(y)\in\mathrm{prox}_{\varphi}(y)$ for any $y\in\mathcal{Y}$);*
(b) *there is a convex C^2 function $\psi:\mathcal{H}\to\mathbb{R}\cup\{+\infty\}$ such that $f(y)=\nabla\psi(y)$ for all $y\in\mathcal{Y}$;*
(c) *the differential $Df(y)$ is a* symmetric positive semi-definite *operator[3] for any $y\in\mathcal{Y}$.*

Corollary 2. [10, Corollary 3] *Let $\mathcal{Y}\subset\mathcal{H}$ be open and convex, and $f:\mathcal{Y}\to\mathcal{H}$ be C^1 with $Df(y)\succ 0$ for any $y\in\mathcal{Y}$. Then f is injective and there is $\varphi:\mathcal{H}\to\mathbb{R}\cup\{+\infty\}$ such that*

[3] A continuous linear operator $L:\mathcal{H}\to\mathcal{H}$ is symmetric if $\langle x,Ly\rangle=\langle Lx,y\rangle$ for any $x,y\in\mathcal{H}$. A symmetric continuous linear operator is positive semi-definite if $\langle x,Lx\rangle\geqslant 0$ for any $x\in\mathcal{H}$. This is denoted $L\succeq 0$. It is positive definite if $\langle x,Lx\rangle>0$ for any nonzero $x\in\mathcal{H}$. This is denoted $L\succ 0$.

$\mathrm{prox}_{\varphi}(y) = \{f(y)\},\ \forall\, y \in \mathcal{Y}$ *and* $\mathrm{dom}(\varphi) = \mathrm{Im}(f)$. *Moreover, if* $x \in \mathcal{H}$ *is a stationary point*[4] *of* $x \mapsto \frac{1}{2}\|y - x\|^2 + \varphi(x)$ *then* $x = f(y)$.

2.2. **Reminders on conditional expectation.** Consider a pair of random variables (X, Y) with values in $\mathbb{R}^n \times \mathbb{R}^m$, with joint probability density function (pdf) $p_{X,Y}(x, y)$ and marginals $p_Y(y), p_X(x)$. For y such that $p_Y(y) > 0$, the conditional distribution of X given $Y = y$ is $p_{X|Y}(x|y) = p_{X,Y}(x, y)/p_Y(y)$. When $\|x\|_2$ is integrable with respect to $p_{X|Y}(\cdot|y)$, the conditional expectation of X given $Y = y$ is

$$\mathbb{E}(X|Y = y) = \int x\ p_{X|Y}(x|y)dx$$

By Bayes rule, the conditional distribution and the marginal $p_Y(y)$ satisfy

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)} \quad \text{and} \quad p_Y(y) = \int p_{X,Y}(x, y)dxdy = \int p_{Y|X}(y|x)p_X(x)dx.$$

Denoting

$$q(y \,|\, x) := p_{Y|X}(y|x)$$

we thus have

$$\mathbb{E}(X|Y = y) \quad = \quad \frac{\int x\ p_{Y|X}(y|x)p_X(x)dx}{p_Y(y)} = \frac{\mathbb{E}_X\ \big(Xq(y \,|\, X)\big)}{\mathbb{E}_X\ \big(q(y \,|\, X)\big)}.$$

As we will see (Corollary 1), the conditional mean has the same expression in related settings such as scalar Poisson denoising. Considering a function $q(y \,|\, x)$ that plays the role of a "proto" conditional distribution of the observation y given the unknown x, we can define "proto" conditional expectation functions in order to characterize when the conditional expectation implements a proximity operator.

2.3. **"Proto" conditional distributions and "proto" conditional expectations.**

Definition 2. *Consider a Hilbert space* $\mathcal{H}$, $\mathcal{X}, \mathcal{Y} \subset \mathcal{H}$ *and a "proto" conditional distribution* $q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+, (x, y) \mapsto q(y \,|\, x)$. *Given a random variable* X *with distribution* P *on* $\mathcal{X}$ *such that*

$$\mathbb{E}_{X\sim P}\ \big((1 + \|X\|)q(y \,|\, X)\big) < \infty, \quad \forall\, y \in \mathcal{Y}. \tag{17}$$

The "proto" marginal distribution $q_P(y)$ *of* $q(y \,|\, x)$ *is defined*[5] *as the function*

$$y \in \mathcal{Y} \mapsto q_P(y) := \mathbb{E}_{X\sim P}\ \big(q(y \,|\, X)\big).$$

On its support

$$\mathcal{Y}_P := \{y \in \mathcal{Y} :\ q_P(y) > 0\},$$

the "proto" conditional mean $f_P(y)$ *is defined as the function*

$$y \in \mathcal{Y}_P \mapsto f_P(y) := \frac{\mathbb{E}_{X\sim P}\ \big(Xq(y \,|\, X)\big)}{\mathbb{E}_{X\sim P}\ \big(q(y \,|\, X)\big)}. \tag{18}$$

[4] u is a stationary point of $\varrho : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ if $\nabla\varrho(u) = 0$; then, ϱ is proper on a neighborhood of u.

[5] If $\mathcal{H} = \mathbb{R}^n$ and $p_{Y|X}(y|x) = q(y \,|\, x)$ is a well-defined conditional probability then $q_P(y) = \int p_{X,Y}(x, y)dx$ is simply the marginal distribution of the random variable Y where $p_{X,Y}(x, y) = p_{Y|X}(y|x)P(x) = q(y \,|\, x)P(x)$.

2.4. **Scalar denoising.** In the scalar case, we fully characterize proto conditional distributions $q(y\,|\,x)$ such that the conditional mean estimator f_P defined by (18) is a proximity operator.

LEMMA 1. *Consider $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$ and $q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$. For P a probability distribution on $\mathcal{X}$ such that* (17) *holds, let f_P, $\mathcal{Y}_P$ be defined as in Definition 2. The following properties are equivalent:*

(a) *For any P satisfying* (17)*, f_P implements a proximity operator;*
(b) *For any $x, x' \in \mathcal{X}, y, y' \in \mathcal{Y}$ with $x' > x$ and $y' > y$, $q(y'\,|\,x')q(y\,|\,x) - q(y'\,|\,x)q(y\,|\,x') \geqslant 0$.*

If $q(y\,|\,x) > 0$ on $\mathcal{X} \times \mathcal{Y}$ then (a)-(b) *are further equivalent to:*

(c) *For any $y, y' \in \mathcal{Y}$ with $y' > y$, the function $x \mapsto \log q(y'\,|\,x) - \log q(y\,|\,x)$ is non-decreasing;*
(d) *For any $x, x' \in \mathcal{Y}$ with $x' > x$, the function $y \mapsto \log q(y\,|\,x') - \log q(y\,|\,x)$ is non-decreasing.*

If $\frac{\partial}{\partial y} \log q(y\,|\,x)$ exists on $\mathcal{X} \times \mathcal{Y}$, then (a)-(b) *are further equivalent to:*

(e) *For any $y \in \mathcal{Y}$, the function $x \mapsto \frac{\partial}{\partial y} \log q(y\,|\,x)$ is non-decreasing.*

If $\frac{\partial}{\partial x} \log q(y\,|\,x)$ exists on $\mathcal{X} \times \mathcal{Y}$, then (a)-(b) *are further equivalent to:*

(f) *For any $x \in \mathcal{X}$ the function $y \mapsto \frac{\partial}{\partial x} \log q(y\,|\,x)$ is non-decreasing.*

The proof is postponed to Annex A.1. Two applications are scalar Poisson denoising (Proposition 1) and denoising in the presence of additive noise (Proposition 2).

Proof. [Proof of Proposition 1] Consider $\mathcal{X} := \mathbb{R}_+^*$, $\mathcal{Y} := \mathbb{R}_+$, and the proto-conditional distribution $q(y\,|\,x) := \frac{x^y}{\Gamma(y+1)} e^{-x}$ which is defined and strictly positive on $\mathcal{X} \times \mathcal{Y}$. For any $y \in \mathcal{Y}$ we have $\sup_{x \in \mathcal{X}} (1 + |x|)\ q(y|x) < \infty$, hence property (17) holds for any distribution P. Definition 2 yields $q_P(y) > 0$ on $\mathcal{Y}$ hence $\mathcal{Y}_P = \mathcal{Y}$. Setting $P = p_X$, as $p_{Y|X}(Y = y|x) = q(y\,|\,x)$ for any $y \in \mathcal{Y}' := \mathbb{N}$ we get $\mathbb{E}\ (X|Y = y) = f_P(y)$ with f_P given by (18). On $\mathcal{X} \times \mathcal{Y}$ we have $\log q(y\,|\,x) = y \log x - \log \Gamma(y + 1) - x$, hence $y \mapsto \frac{\partial}{\partial x} \log q(y\,|\,x) = \frac{y}{x} - 1$ is non-decreasing on $\mathcal{X}$. The fact that f_P implements a proximity operator, i.e., the existence of $\tilde{\varphi}_X$, follows by Lemma 1(a)⇔(e). □

Proof. [Proof of Proposition 2] Consider $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, $q(y\,|\,x) := p_{Y|X}(y|x) = \exp(-F(y - x))$, and observe that $\log q(y\,|\,x) = -F(y - x)$. For $P := p_X$ a probability distribution satisfying (17), reasoning as in the proof of Proposition 1 we have $q_P(y) > 0$ on $\mathcal{Y}$ hence $\mathcal{Y}_P = \mathcal{Y}$ and $\mathbb{E}(X|Y = y) = f_P(y)$.

Consider first the case where F is C^1. Then, F is convex iff $u \mapsto F'(u)$ is non-increasing, which is equivalent to $y \mapsto \frac{\partial}{\partial x} \log q(y\,|\,x) = F'(y-x)$ being non-decreasing. By Lemma 1(f)⇔(a) this is equivalent to the fact that f_P implements a proximity operator for any P satisfying (17). To conclude, we exploit a simple lemma proved in Appendix A.7.

LEMMA 2. *If $G : \mathbb{R} \to \mathbb{R}$ is convex and satisfies $\int_{\mathbb{R}} e^{-G(x)} dx < \infty$ then $\sup_{x \in \mathbb{R}} (1+|x|) e^{-G(x)} < \infty$.*

Hence:

- if F is not convex then there exists P satisfying (17) such that f_P does *not* implement a proximity operator, hence Proposition 2(b) cannot hold;
- if F is convex then for any y the function $G : x \mapsto F(y-x)$ is convex and $\int_{\mathbb{R}} e^{-G(x)} dx < \infty$ hence, by Lemma 2, the function $x \mapsto (1+|x|) q(y\,|\,x)$ is bounded. As a result, (17) holds for any distribution P. As just shown, f_P implements a proximity operator *for any* P.

This establishes the equivalence Proposition 2(a)⇔(b) when F is C^1.

To extend the result when F is only C^0, we reason similarly using Lemma 1(d)⇔(a). A bit more work is needed to show that Lemma 1(d) holds iff F is convex, as we now establish.

With the change of variable $u = y - x'$, $h = x' - x > 0$, Lemma 1(d) is equivalent to:

$$\forall\, h > 0, \quad u \mapsto F(u+h) - F(u) \quad \text{is non-decreasing.} \tag{19}$$

When (19) holds we have for any $u_1 < u_2$ (using $h := (u_2 - u_1)/2$ in (19))

$$F(\tfrac{u_1+u_2}{2}) - F(u_1) = F(u_1+h) - F(u_1) \leqslant F(\tfrac{u_1+u_2}{2} + h) - F(\tfrac{u_1+u_2}{2}) = F(u_2) - F(\tfrac{u_1+u_2}{2})$$

hence $F(\frac{u_1+u_2}{2}) \leqslant \frac{F(u_1)+F(u_2)}{2}$. As F is C^0, this is well known to imply that F is convex.

Vice-versa, when F is convex, given u and $h, h' > 0$ we wish to prove that $F(u_4) - F(u_3) \geqslant F(u_2) - F(u_1)$ where $u_1 := u$, $u_2 := u+h$,$u_3 := u+h'$, $u_4 := u+h+h'$. Two cases are possible: either $u_1 < u_2 < u_3 < u_4$ or $u_1 < u_3 \leqslant u_2 < u_4$. We treat the latter, the first one can be handled similarly. Since F is convex and $u_3 \leqslant u_2$, there exists $a \in \partial F(u_3)$ and $b \in \partial F(u_2)$ with $a \leqslant b$, and we get $a(u_3 - u_1) = ah' \leqslant bh' = b(u_4 - u_2)$. As a result

$$\begin{aligned} F(u_4) - F(u_3) &= F(u_4) - F(u_2) + F(u_2) - F(u_3) \geqslant b(u_4 - u_2) + F(u_2) - F(u_3) \\ &\geqslant a(u_3 - u_1) + F(u_2) - F(u_3) \\ &\geqslant F(u_3) - F(u_1) + F(u_2) - F(u_3) = F(u_2) - F(u_1). \end{aligned}$$

□

Example 2. [Completely separable model] Consider a completely separable model: the entries $X_i \in \mathbb{R}$, $1 \leqslant i \leqslant n$ of the random variable $X \in \mathbb{R}^n$ are drawn independently with prior distributions P_i, i.e., $p_X(x) = \prod_{i=1}^n p_{X_i}(x_i)$; the conditional distribution of $Y \in \mathbb{R}^n$ given $x \in \mathbb{R}^n$ corresponds to independent noise on each coordinate, $p_{Y|X}(y|x) = \prod_{i=1}^n p_{Y_i|X_i}(y_i|x_i)$. This implies that the conditional expectation can be computed coordinate by coordinate. Assuming further that $p_{Y_i|X_i}(y_i|x_i) \propto q_i(y_i|x_i)$ where q_i satisfies Lemma 1-(b), we obtain for $1 \leqslant i \leqslant n$

$$\mathbb{E}(X|Y=y)_i = \mathbb{E}(X_i|Y_i=y_i) \;\in\; \arg\min_{x_i \in \mathbb{R}} \left\{ \tfrac{1}{2}(y_i - x_i)^2 + \varphi_{P_i}(x_i) \right\}, \quad \forall\, y \in \mathbb{R}.$$

and as a result

$$\mathbb{E}(X|Y=y) \;\in\; \arg\min_{x \in \mathbb{R}^n} \left\{ \tfrac{1}{2}\|y - x\|^2 + \sum_{i=1}^n \varphi_{P_i}(x) \right\}, \quad \forall\, y \in \mathbb{R}^n.$$

Hence, the conditional mean implements the proximity operator of $\varphi_P(x) := \sum_{i=1}^n \varphi_{P_i}(x_i)$.

2.5. **Multivariate denoising.** In dimension $1 \leqslant n < \infty$, we have the following result.

Lemma 3. *Consider $\mathcal{X}, \mathcal{Y} \subset \mathcal{H}$ and $q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$. For P a probability distribution on $\mathcal{X}$ such that* (17) *holds, let f_P, $\mathcal{Y}_P$ be defined as in Definition 2.*

Assume that for any P satisfying (17)*, f_P implements a proximity operator. Then, for any $x, x' \in \mathcal{X}$, $y, y' \in \mathcal{Y}$:*

(a) *if $\nabla_x \log q(y \,|\, x)$ and $\nabla_x \log q(y' \,|\, x)$ exist, there is a scalar $c = c(x, y, y') \geqslant 0$ such that*

$$\nabla_x \log q(y' \,|\, x) - \nabla_x \log q(y \,|\, x) = c\ (y' - y). \tag{20}$$

(b) *if* $\nabla_y \log q(y \,|\, x)$ *and* $\nabla_y \log q(y \,|\, x')$ *exist, there is a scalar* $c = c(x, x', y) \geqslant 0$ *such that*

$$\nabla_y \log q(y \,|\, x') - \nabla_y \log q(y \,|\, x) = c' \ (x' - x). \tag{21}$$

The proof is postponed to Annex A.2.

REMARK 5. *In* (a)-(b) *the gradients are only assumed to exist at particular points* x, x', y, y'*, leading to the necessary conditions* (20)-(21) *at these points.*

A first consequence is that MMSE estimation with additive Laplacian noise (resp. with Poisson noise) in dimension $n \geqslant 2$ behaves differently from dimension $n = 1$ (cf Proposition 1, Proposition 2).

EXAMPLE 3 (multivariate additive Laplacian noise). *Consider a multivariate Laplacian noise model: given* $x \in \mathcal{X} = \mathbb{R}^n$*, the conditional probability of the random vector* Y *on* $\mathcal{Y} = \mathbb{R}^n$ *is defined by* $p(y|x) \propto q(y \,|\, x) := e^{-\|x-y\|_1}$*. As* $\log q(y \,|\, x) = -\|x - y\|_1$*, given* $y = (y_i)_{i=1^n}$ *and* $x = (x_i)_{i=1}^n$ *such that* $x_i \neq y_i$*,* $\forall\, 1 \leqslant i \leqslant n$*,* $\log q(\cdot, y)$ *is differentiable at* x *and* $\nabla_x \log q(y \,|\, x) = -\mathrm{sign}(x - y)$*. Hence, for* $x \in \mathcal{X}$*,* $y, y' \in \mathcal{Y}$ *such that* $x_i \notin \{y_i, y_i'\}$ *for* $1 \leqslant i \leqslant n$ *we have*

$$\nabla_x \log q(y' \,|\, x) - \nabla_x \log q(y \,|\, x) = -\big(\mathrm{sign}(x - y') - \mathrm{sign}(x - y)\big). \tag{22}$$

- *The scalar case* $n = 1$ *is covered by Proposition 2 since* q *is log-concave. For any distribution* P *on the scalar random variable* $X \in \mathbb{R}$*, the MMSE estimator* f_P *implements a proximity operator.*
- *For* $n \geqslant 2$*, this is no longer the case. Consider* $x = 0$ *and* $y \in (\mathbb{R}_+^*)^n$ *such that* $y_1 \neq y_2$*, and* $y' = -y$*. The vector* $-(\mathrm{sign}(x - y') - \mathrm{sign}(x - y)) = 2 \cdot \mathbf{1}_n$ *is not proportional to the vector* $(y' - y) = 2y'$ *which first two entries are distinct. Hence* (22) *is incompatible with condition* (20) *and by Lemma 3, there exists a prior distribution* P *such that* f_P *does* not *implement a proximity operator.*

REMARK 6. *As the noise distribution is separable (multivariate Laplacian noise corresponds to i.i.d. scalar noise on each coordinate), by Example 2 the MMSE estimator in fact implements a proximity operator as soon as the prior* P *is also separable (i.e., if* X *has independent entries* X_i*). However, for* $n \geqslant 2$*, we have just shown that there exists a prior* P *(non-separable) such that* f_P *does not implement a proximity operator.*

EXAMPLE 4 (multivariate Poisson denoising). *Consider a multivariate Poisson noise model: given* $x \in \mathcal{X} := (\mathbb{R}_+^*)^n$*, the conditional probability of the random vector of integers* Y *on* $\mathcal{Y} := \mathbb{N}^n$ *is defined by* $p(y|x) = q(y \,|\, x) := \prod_{i=1}^n \left(\frac{x_i^{y_i} e^{-x_i}}{\Gamma(y_i+1)}\right)$*. We observe that*

$$\log q(y \,|\, x) = \sum_{i=1}^n y_i \ln x_i - x_i - \log \Gamma(y_i + 1)$$

Given $y \in \mathcal{Y}$*, the function* $x \mapsto \log q(y \,|\, x)$ *is differentiable on* $\mathcal{X}$ *with* $\nabla_x \log q(y \,|\, x) = (y_i/x_i - 1)_{i=1}^n$ *hence for* $x \in \mathcal{X}$*,* $y, y' \in \mathcal{Y}$ *we have*

$$\nabla_x \log q(y' \,|\, x) - \nabla_x \log q(y \,|\, x) = ((y_i' - y_i)/x_i)_{i=1}^n.$$

- *The case* $n = 1$ *is covered by Proposition 1:* f_P *implements a proximity operator for any prior* P*.*

- *For $n \geqslant 2$ this is again no longer the case. Consider, e.g., $x \in \mathcal{X}$ with $x_1 \neq x_2$ and $y, y' \in \mathcal{Y}$ such that $y_i \neq y'_i$, $i = 1, 2$. The vector with entries $(y'_i - y_i)/x_i$ cannot be proportional to the vector $y' - y$ hence* (20) *cannot hold. By Lemma 3, there is a prior P such that f_P does* not *implement a proximity operator.*

Remark 6 again applies: for a separable prior P, f_P implements a proximity operator, yet there exists a (non-separable) prior P such that f_P does not implement a proximity operator.

For smooth enough proto-conditional distributions we have the following corollary of Lemma 3.

LEMMA 4. *Let $\mathcal{H}$ be of dimension $2 \leqslant n \leqslant \infty$ and consider open sets $\mathcal{X}, \mathcal{Y} \subset \mathcal{H}$ where $\mathcal{X}$ is connected. Let $q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+^*$ be such that $x \mapsto \nabla_x \log q(y \,|\, x)$ is C^0 for any $y \in \mathcal{Y}$, and $y \mapsto \nabla_y \log q(y \,|\, x)$ is C^0 for any $x \in \mathcal{X}$. For P a probability distribution on $\mathcal{X}$ satisfying* (17)*, let f_P, $\mathcal{Y}_P$ be defined as in Definition 2.*

If f_P implements a proximity operator for any such P, then there exists $c \geqslant 0$, $a \in C^1(\mathcal{X})$ $b \in C^1(\mathcal{Y})$ such that

$$q(y \,|\, x) = \exp\left(-a(x) - b(y) + c\langle x, y\rangle\right), \quad \forall\, x, y \in \mathcal{X} \times \mathcal{Y}. \tag{23}$$

The proof is postponed to Annex A.3. A converse result also holds.

LEMMA 5. *Consider $\mathcal{X}, \mathcal{Y} \subset \mathcal{H}$ where $\mathcal{Y}$ is open and convex. Assume that $q : \mathcal{X} \times Y \to \mathbb{R}_+^*$ has the expression* (23) *where $c \geqslant 0$, $a : \mathcal{X} \to \mathbb{R}$, $b : \mathcal{Y} \to \mathbb{R}$, b is $C^1(\mathcal{Y})$, and b' is locally bounded*[6]*.*

Let P be a probability distribution satisfying: for any $y \in \mathcal{Y}$, there is $r = r(y) > 0$ such that

$$\mathbb{E}_{X \sim P}\left\{(1 + \|X\|)^2 \cdot \exp\left(-a(X) + cr\|X\| + c\langle X, y\rangle\right)\right\} < \infty. \tag{24}$$

Then (17) *holds and, using the notations of Definition 2, we have*

- *the proto-conditional mean f_P is differentiable on $\mathcal{Y}_P = \mathcal{Y}$;*
- *its differential is symmetric positive semi-definite*[7]*, i.e., $Df_P(y) \succeq 0, \forall\, y \in \mathcal{Y}$;*
- *the proto-conditional mean f_P implements the proximity operator of a penalty φ_P.*

Moreover if $c > 0$ and if the support of the distribution P is not included in any hyperplane (i.e. if $P(\langle X, u\rangle = d) < 1$ for any nonzero $u \in \mathcal{H}$ and any $d \in \mathbb{R}$), then φ_P can be chosen such that:

- *$Df_P(y) \succ 0$ for any y;*
- *the proto-conditional mean f_P is injective;*
- *for any y, $f_P(y)$ is the unique stationary point (and global minimizer) of $x \mapsto \frac{1}{2}\|y - x\|^2 + \varphi_P(x)$. In particular, f_P* is *the proximity operator of φ_P.*

The proof is postponed to Appendix A.4.

REMARK 7. *The case $c = 0$ corresponds to independent random variables X and Y governed by $p_{Y|X}(y|x) = \frac{q(y \,|\, x)}{\int q(y' \,|\, x) dy'} = \frac{\exp(-a(x) - b(y))}{\int \exp(-a(x) - b(y')) dy'} = \frac{\exp(-b(y))}{\int \exp(-b(y')) dy'}$.*

A consequence of Lemma 5 is that we recover the results of [7]. We also cover a variant of multivariate Poisson denoising that leads again to a proximity operator even for $n \geqslant 2$ as expressed in Proposition 5.

[6] If $\mathcal{H}$ is finite dimensional then local boundedness is automatic by compactness arguments; the assumption is useful in infinite dimension.

[7] see footnote 3 page 10.

2.5.1. *Proof of Proposition 3.* Consider white Gaussian noise, i.e. $p(y|x) \propto q(y\,|\,x)$ with $q(y\,|\,x) := \exp\left(-\frac{c}{2}\|x-y\|^2\right)$; $\mathcal{X} = \mathcal{Y} = \mathcal{H} = \mathbb{R}^n$. Observe that $\log q(y\,|\,x) = -\frac{c}{2}\|x-y\|^2 = c\langle x, y\rangle - a(x) - b(y)$ with $a(x) := \frac{c}{2}\|x\|^2$, $b(y) := \frac{c}{2}\|y\|^2$. Since $\sup_{x\in\mathcal{X}}(1+\|x\|)^2 \exp\left(-a(x) + (r + c\|y\|)\|x\|\right) \leqslant \sup_{u\in\mathbb{R}_+}(1+u)^2 e^{-cu^2/2 + (r+c\|y\|)u} < \infty$, any probability distribution P satisfies (24) hence we can apply Lemma 5.

2.5.2. *Proof of Proposition 5.* Consider $\mathcal{X} := (\mathbb{R}_+^*)^n$, $\mathcal{Y}' = \mathbb{N}^n$, $\mathcal{Y} := (\mathbb{R}_+)^n$. For $(x, y) \in \mathcal{X} \times \mathcal{Y}'$ we have $p(y|x) = q(y\,|\,x)$ where

$$\log q(y\,|\,x) = \sum_{i=1}^{n} \left(y_i \log x_i - x_i - \log \Gamma(y_i + 1)\right)$$

is defined on $\mathcal{X} \times \mathcal{Y}$. Denoting $z := (\log x_i)_{i=1}^n \in \mathcal{Z} := \mathbb{R}^n$ we have $\widetilde{p}(y|z) := \widetilde{q}(y|z)$ where $\log \widetilde{q}(y|z) := \langle z, y\rangle - a(z) - b(y)$ with

$$a(z) := \sum_{i=1}^{n} e^{z_i}$$

and $b(y) := \sum_{i=1}^n \log\Gamma(y_i + 1)$.

For any $y \in \mathcal{Y}$, $r > 0$, using arguments similar to those in the proof of Lemma 2, we get that

$$\sup_{z\in\mathcal{Z}} (1 + \|z\|)^2 \exp\left(-a(z) + (r + \|y\|)\|z\|\right) < \infty$$

we get that for any distribution p_X on X, denoting P the resulting distribution on the random variable $Z = \log X \in \mathcal{Z}$, P necessarily satisfies (24) for any $y \in \mathcal{Y}$ and $r > 0$.

As b is $C^1(\mathcal{Y})$, both b and b' are locally bounded hence, we can again apply Lemma 5 to get $f_P(y) = \mathbb{E}_{Z\sim P}(Z|Y = y) = \mathbb{E}(\log X|Y = y)$ implements a proximity operator.

Appendix A. Proofs

A.1. **Proof of Lemma 1.** (a)⇒(b). First, we establish a necessary condition that any function $q(x\,|\,y)$ must satisfy to ensure that (18) implements a proximity operator for any prior probability distribution P on the random variable X. This condition will be re-used for the proof of Lemma 3.

Lemma 6. *Consider $\mathcal{X}, \mathcal{Y} \subset \mathcal{H}$ and $q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+, (x, y) \mapsto q(y\,|\,x)$. Assume that* for any probability distribution P such that (17) holds*, f_P implements the proximity operator of some penalty φ_P (i.e. $f_P(y) \in \mathrm{prox}_{\varphi_P}(y)$, $\forall\, y \in \mathcal{Y}_P$), with $\mathcal{Y}_P$ and f_P as in Definition 2. Then the function q satisfies*

$$\forall\, x, x' \in \mathcal{X}, y, y' \in \mathcal{Y}, \qquad \langle x' - x, y' - y\rangle \left(q(y'\,|\,x')q(y\,|\,x) - q(y'\,|\,x)q(y\,|\,x')\right) \geqslant 0. \tag{25}$$

Even though (25) is not intuitive, its main interest is that it depends only on the "proto" conditional distribution $q(y\,|\,x)$ *but not on the prior distribution P on X*. It necessarily holds when the "proto" conditional distribution $q(y\,|\,x)$ is such that *for any prior probability distribution P on X*, the "proto" conditional expectation f_P implements a proximity operator.

Proof. Given $x, x' \in \mathcal{X}$, $y, y' \in \mathcal{Y}$, consider the probability distribution $P := \frac{1}{2}(\delta_x + \delta_{x'})$. It is straightforward that P satisfies (17).

Now, denoting $\lambda := (q(y\,|\,x)+q(y\,|\,x'))(q(y'\,|\,x)+q(y'\,|\,x'))$, we distinguish two cases depending whether $\lambda = 0$ or not.

First, if $\lambda = 0$ then one must have $q(y\,|\,x) + q(y\,|\,x') = 0$, or $q(y'\,|\,x) + q(y'\,|\,x') = 0$, or both. Without loss of generality we treat the case $q(y\,|\,x) + q(y\,|\,x') = 0$ (the other cases are treated similarly). Since q is non-negative, this implies $q(y\,|\,x) = q(y\,|\,x') = 0$, hence $q(y'\,|\,x')q(y\,|\,x) - q(y'\,|\,x)q(y\,|\,x') = 0$ and (25) trivially holds.

Now, consider the case $\lambda > 0$, i.e. $q(y\,|\,x) + q(y\,|\,x') > 0$ and $q(y'\,|\,x) + q(y'\,|\,x') > 0$. By the definition of q_P and $\mathcal{Y}_P$ in Definition 2, this implies $y, y' \in \mathcal{Y}_P$ and thus $q_P(y) > 0$ and $q_P(y') > 0$. Hence, by the assumption of Lemma 6, f_P implements a proximity operator, therefore by Theorem 1 there exists a convex lsc function ψ_P such that $f_P(y) \in \partial\psi_P(y)$ and $f_P(y') \in \partial\psi_P(y')$. From the definition of a subdifferential, this implies that $\psi_P(y')-\psi_P(y) \geqslant \langle f_P(y), y'-y\rangle$ and that $\psi_P(y) - \psi_P(y') \geqslant \langle f_P(y'), y - y'\rangle$. Adding both inequalities yields[8] $0 \geqslant \langle f_P(y) - f_P(y'), y' - y\rangle$. Since by (18)

$$f_P(y) = \frac{xq(y\,|\,x) + x'q(y\,|\,x')}{q(y\,|\,x) + q(y\,|\,x')} \quad \text{and} \quad f_P(y') = \frac{xq(y'\,|\,x) + x'q(y'\,|\,x')}{q(y'\,|\,x) + q(y'\,|\,x')}$$

we obtain with straightforward computations that

$$0 \leqslant \lambda\langle f_P(y') - f_P(y), y' - y\rangle = \langle x' - x, y' - y\rangle \left(q(y'\,|\,x')q(y\,|\,x) - q(y'\,|\,x)q(y\,|\,x')\right).$$

□

In the scalar case (25) is equivalent to Lemma 1(b), hence Lemma 6 establishes (a)⇒(b).

(b)⇒(a). Since we consider the scalar case, the assumption that (b) holds implies (25). Consider $y, y' \in \mathcal{Y}_P$ and X, X' two independent random variables with distribution P. Write

$$\begin{aligned}
q_P(y)q_P(y')\,\big(f_P(y') - f_P(y)\big) &\overset{(18)}{=} q_P(y)\,\mathbb{E}_{X'}\left(X'\,q(y'\,|\,X')\right) - q_P(y')\,\mathbb{E}_{X'}\left(X'\,q(y\,|\,X')\right) \\
&= \mathbb{E}_X\big(q(y\,|\,X)\big)\,\mathbb{E}_{X'}\left(X'\,q(y'\,|\,X')\right) \\
&\quad -\mathbb{E}_X\left(q(y'\,|\,X)\right)\,\mathbb{E}_{X'}\left(X'\,q(y\,|\,X')\right) \\
&= \mathbb{E}_{X,X'}\Big(X'\left(q(y\,|\,X)q(y'\,|\,X') - q(y'\,|\,X)q(y\,|\,X')\right)\Big) \qquad (26) \\
&\overset{X\leftrightarrow X'}{=} \mathbb{E}_{X',X}\Big(X\left(q(y\,|\,X')q(y'\,|\,X) - q(y'\,|\,X')q(y\,|\,X)\right)\Big) \qquad (27) \\
&\overset{((26)+(27))/2}{=} \mathbb{E}_{X',X}\Big(\tfrac{X'-X}{2}\left(q(y'\,|\,X')q(y\,|\,X) - q(y'\,|\,X)q(y\,|\,X')\right)\Big)
\end{aligned}$$

If follows that

$$\begin{aligned}
(y'-y)\big(f_P(y') - f_P(y)\big) &= \tfrac{1}{2q_P(y)q_P(y')} \cdot \mathbb{E}_{X',X}\left\{\tfrac{(X'-X)(y'-y)}{2}\left(q(y'\,|\,X')q(y\,|\,X) - q(y'\,|\,X)q(y\,|\,X')\right)\right\} \\
&\overset{(25)}{\geqslant} 0,
\end{aligned}$$

hence f_P is non-decreasing on the set $\mathcal{Y}_P$. By Corollary 1, f_P implements a proximity operator.

[8] or equivalently $\langle f_P(y') - f_P(y), y' - y\rangle \geqslant 0$. Since this holds for any $y, y' \in \mathcal{Y}$, f_P is a *monotone* operator (see, e.g., [16]).

(b)⇔(c)⇔(d) when $q(y\,|\,x) > 0$ on $\mathcal{X} \times \mathcal{Y}$. We sketch the proof of the equivalence of (b) and (c). The equivalence between (b) and (d) follows similarly. Denote $Q(x; y, y') := \log q(y'\,|\,x) - \log q(y\,|\,x)$. Property (b) holds if and only if for any $x, x' \in \mathcal{X}$, $y, y' \in \mathcal{Y}$ with $x' > x$, $y' > y$ we have $q(y'\,|\,x')q(y\,|\,x) \geqslant q(y'\,|\,x)q(y\,|\,x')$, which is equivalent to $\log q(y'\,|\,x') + \log q(y\,|\,x) - \log q(y'\,|\,x) - \log q(y\,|\,x') \geqslant 0$, that is to say $Q(x'; y, y') - Q(x; y, y') \geqslant 0$. i.e., $x \mapsto Q(x; y, y')$ is non-decreasing as soon as $y' > y$.

(c)⇔(e) and (d)⇔(f). This is straightforward using, e.g., that $\frac{\partial}{\partial y} \log q(y\,|\,x) = \lim_{y' \to y} \frac{Q(x;y,y')}{y'-y}$.

A.2. **Proof of Lemma 3.** We establish (a).(b) is obtained similarly by reversing the roles of x and y.

Since f_P is a proximity operator for any probability distribution P such that (17) holds, by Lemma 6 it follows that property (25) holds.

Consider $h \in \mathcal{H}$ such that $\langle y' - y, h\rangle > 0$. As $\log q(\cdot, y)$ is differentiable at x, we have $x' := x + \varepsilon h \in \mathcal{X}$ for $\varepsilon > 0$ small enough, and $\langle x' - x, y' - y\rangle = \varepsilon\langle h, y' - y\rangle > 0$. By (25) we have $q(y'\,|\,x')q(y\,|\,x) \geqslant q(y'\,|\,x)q(y\,|\,x')$. Taking the logarithm yields

$$\log q(y'\,|\,x + \varepsilon h) - \log q(y'\,|\,x) \geqslant \log q(y\,|\,x + \varepsilon h) - \log q(y\,|\,x)$$

As $\log q(\cdot, y)$ and $\log q(\cdot, y')$ are both differentiable at x, it follows that

$$\big\langle \nabla_x \log q(y'\,|\,x), \varepsilon h\big\rangle + o(\varepsilon) \geqslant \langle \nabla_x \log q(y\,|\,x), \varepsilon h\rangle + o(\varepsilon)$$

hence $\langle \nabla_x \log q(y'\,|\,x) - \nabla_x \log q(y\,|\,x), h\rangle \geqslant 0$. Since this holds for any h such that $\langle y' - y, h\rangle > 0$, there is a scalar $c \geqslant 0$ (a priori dependent on x, y, y') such that (20) holds.

A.3. **Proof of Lemma 4.** Since $\dim \mathcal{H} \geqslant 2$ and $\mathcal{X}$ and $\mathcal{Y}$ are open, the affine dimension of $\mathcal{X}$ (resp. of $\mathcal{Y}$) exceeds two. We use the following lemma.

Lemma 7. *Let $\mathcal{Z} \subset \mathcal{H}$ be of affine dimension at least two (i.e., there is no pair $z_1, z_2 \in \mathcal{H}$ such that $\mathcal{Z} \subset \{tz_1 + (1-t)z_2,\ t \in \mathbb{R}\}$). Assume that the function $\theta : \mathcal{Z} \to \mathcal{H}$ satisfies*

$$\forall\, z, z' \in \mathcal{Z},\ \exists c(z, z') \in \mathbb{R},\ \theta(z) - \theta(z') = c(z, z')\ (z - z'). \tag{28}$$

Then, there exists $c \in \mathbb{R}$ such that

$$\forall\, z, z' \in \mathcal{Z},\ \theta(z) - \theta(z') = c\ (z - z').$$

Proof. **Step 1.** We show that if $z_1, z_2, z_3 \in \mathcal{Z}$ are not aligned (i.e., if they are affinely independent) then $c(z_i, z_j) = c(z_1, z_2)$ for all $1 \leqslant i \neq j \leqslant 3$.

Denote $c_{ij} := c(z_i, z_j)$. By symmetry(28) yields $c_{ij} = c_{ji}$, $\forall\ i \neq j$. Moreover, summing up (28) with $z = z_i$, $z' = z_j$ over $(i, j) \in \{(1, 2), (2, 3), (3, 1)\}$ yields

$$0 = c_{12}(z_1 - z_2) + c_{23}(z_2 - z_3) + c_{31}(z_3 - z_1) = (c_{12} - c_{31})z_1 + (c_{23} - c_{12})z_2 + (c_{31} - c_{23})z_3.$$

As the coefficients in the right hand side sum to zero, affine independence yields $c_{12} = c_{23} = c_{31}$.

Step 2. As the affine dimension of $\mathcal{Z}$ exceeds two it is not a singleton and we can choose an arbitrary pair $z_1 \neq z_2 \in \mathcal{Z}$. Define $c := c(z_1, z_2)$. We show that for any $x, y \in \mathcal{Z}$ we have $c(x, y) = c(y, x) = c$.

Define $\mathcal{S} := \{z_1 + (1-t)z_2,\ t \in \mathbb{R}\}$ the affine hull of z_1, z_2 and observe that $\mathcal{Z} \cap \mathcal{S}^c \neq \varnothing$.

First, consider $x \in \mathcal{Z} \cap \mathcal{S}^c$. For $y = z_1$, as z_1, z_2, x are not aligned, by Step 1 we have $c(x, z_1) = c(z_1, x) = c$. For $y \in \mathcal{Z} \cap \mathcal{S} \backslash \{z_1\}$, as x, y, z_1 are not aligned, by Step 1 again we get $c(x, y) = c(y, x) = c(x, z_1) = c$. This establishes the result for any $x \in \mathcal{Z} \cap \mathcal{S}^c$ and $y \in \mathcal{Z} \cap \mathcal{S}$.

Second, consider $x, y \in \mathcal{Z} \cap \mathcal{S}^c$. As just shown, we have $c(x, z_1) = c(z_1, y) = c$, hence by (28)

$$\theta(x) - \theta(y) = \theta(x) - \theta(z_1) + \theta(z_1) - \theta(y) = c(x, z_1)\ (x - z_1) + c(z_1, y)\ (z_1 - y) = c\ (x - y).$$

Finally consider $x, y \in \mathcal{Z} \cap \mathcal{S}$. Let $z \in \mathcal{Z} \cap \mathcal{S}^c$ be arbitrary. As $c(x, z) = c(z, y) = c$, (28) yields

$$\theta(x) - \theta(y) = \theta(x) - \theta(z) + \theta(z) - \theta(y) = c(x, z)\ (x - z) + c(z, y)\ (z - y) = c(x - y).$$

□

For a given $x \in \mathcal{X}$, consider the function $y \mapsto \theta_x(y) := \nabla_x \log q(y \,|\, x)$. By Lemma 3(a), θ_x satisfies (28) and the constants $c(z, z')$ are non-negative. By Lemma 7 there is $c_x \in \mathbb{R}_+$ such that

$$\theta_x(y) - \theta_x(y') = c_x\ (y - y'),\ \forall\ y, y' \in \mathcal{Y}. \tag{29}$$

As a result $y \mapsto \theta_x(y)$ is differentiable with $D_y \nabla_x \log q(y \,|\, x) = c_x$ Id. Similarly, given $y \in \mathcal{Y}$, with $\varrho_y(x) := \nabla_y \log q(y \,|\, x)$, there is $d_y \in \mathbb{R}_+$ such that

$$\varrho_y(x) - \varrho_y(x') = d_y\ (x - x'),\ \forall\ x, x' \in \mathcal{X} \tag{30}$$

hence $D_x \nabla_y \log q(y \,|\, x) = d_y$ Id.

When $(x, y) \mapsto \log q(y \,|\, x)$ is C^2, by Schwarz' theorem we have $D_x \nabla_y \log q(y \,|\, x) = D_y \nabla_x \log q(y \,|\, x)$ for any $x, y \in \mathcal{X} \times \mathcal{Y}$. Thus, c_x Id $= d_y$ Id for any $x, y \in \mathcal{X} \times \mathcal{Y}$ hence $c_x = d_y = c \geqslant 0$ is independent of x, y and

$$\nabla_x \log q(y \,|\, x) - \nabla_x \log q(y' \,|\, x) = c(y - y'), \quad \forall\ x \in \mathcal{X}, \forall\ y, y' \in \mathcal{Y}. \tag{31}$$

Let us now show that (31) also holds with the considered weaker assumption on q . For this, fix some arbitrary $x \in \mathcal{X}$ and $x' \in \mathcal{X}$ close enough so that $\{x + t(x' - x)\} \subset \mathcal{X}$ (remember that $\mathcal{X}$ is open so this is possible). Consider $y, y' \in \mathcal{Y}$. Denote

$$f(t) := \log q(y \,|\, x + t(x' - x)) - \log q(y' \,|\, x + t(x' - x)),\ t \in [0, 1].$$

As $x \mapsto \nabla_x \log q(y \,|\, x)$ is assumed to be continuous, the function f is C^1 on $[0, 1]$ and by (29) we have

$$f'(t) = \langle \nabla_x \log q(y \,|\, x + t(x' - x)) - \nabla_x \log q(y' \,|\, x + t(x' - x)), x' - x \rangle = c_{x + t(x' - x)} \langle y - y', x' - x \rangle.$$

As a result

$$\begin{aligned} \log q(y \,|\, x') - \log q(y' \,|\, x') - \log q(y \,|\, x) + \log q(y' \,|\, x) &= f(1) - f(0) = \int_0^1 f'(t) dt \\ &= \int_0^1 c_{x + t(x' - x)} dt\ \langle y - y', x' - x \rangle., \end{aligned}$$

By (30), taking the gradient of both sides with respect to y yields

$$d_y\ (x' - x) = \int_0^1 c_{x + t(x' - x)} dt\ (x' - x).$$

Thus, d_y does not depend on y. Similarly, c_x does not depend on x. This establishes (30).

Fix an arbitrary $y_0 \in \mathcal{Y}$ and denote $H(x,y) := \log q(y\,|\,x) - c\langle x, y\rangle$ and $a(x) := -H(x, y_0)$. We obtain

$$\nabla_x \left(H(x,y) + a(x)\right) = \nabla_x \left(H(x,y) - H(x,y_0)\right) = 0, \quad \forall\, x \in \mathcal{X}, \forall\, y \in \mathcal{Y}.$$

Since $\mathcal{X}$ is open and connected, hence path-connected, it follows that for any $y \in \mathcal{Y}$ there is $b(y) \in \mathbb{R}$ such that: $H(x,y) + a(x) = -b(y)$ for all $x \in \mathcal{X}$, i.e.,

$$\log q(y\,|\,x) = -a(x) - b(y) + c\langle x, y\rangle.$$

As both $x \mapsto \nabla_y \log q(y\,|\,x)$ and $y \mapsto \nabla_x \log q(y\,|\,x)$ are C^1, both $a(\cdot)$ and $b(\cdot)$ are C^1.

A.4. **Proof of Lemma 5.** We use the following lemma, which proof is slightly postponed.

LEMMA 8. *Consider $\mathcal{H}$ a Hilbert space, two subsets $\mathcal{X}, \mathcal{Y} \subset \mathcal{H}$ and a "proto" conditional distribution $q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$. Let $\mathcal{V} \subset \mathcal{Y}$ be an open set such that the gradient $\nabla_y q(y\,|\,x)$ exists for any $x \in \mathcal{X}$, $y \in \mathcal{V}$. Define*

$$C_{q,\mathcal{V}}(x) := \sup_{y \in \mathcal{V}} \left\{ q(y\,|\,x) + \|\nabla_y q(y\,|\,x)\|_{\mathcal{H}} \right\} \in [0, \infty] \tag{32}$$

Consider P a probability distribution such that

$$\mathbb{E}_{X \sim P} \left\{ (1 + \|X\|_{\mathcal{H}})\; C_{q,\mathcal{V}}(X) \right\} < \infty. \tag{33}$$

Then (17) holds, the set $\mathcal{V}_P := \{y \in \mathcal{V} :\; q_P(y) > 0\} = \mathcal{V} \cap \mathcal{Y}_P$ is open, and

- *the "proto" marginal distribution $q_P(y) := \mathbb{E}_{X \sim P}\; (q(y\,|\,X))$ is differentiable on $\mathcal{V}$;*
- *the "proto" conditional mean f_P defined by (18) is differentiable on $\mathcal{V}_P$ with*[9]

(34)

$$Df_P(y) = \tfrac{1}{2q_P^2(y)}\; \mathbb{E}_{X \sim P} \mathbb{E}_{X' \sim P} \Big(q(y\,|\,X) q(y\,|\,X') \cdot (X' - X) \left(\nabla_y^T \log q(y\,|\,X') - \nabla_y^T \log q(y\,|\,X) \right) \Big)$$

where by convention $\nabla_y \log q(y\,|\,x) = 0$ when $q(y\,|\,x) = 0$.

REMARK 8. *The assumption (33) is chosen for simplicity but could be relaxed.*

By (23) and the fact that $b \in C^1(\mathcal{Y})$ we have for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$

$$\begin{aligned} q(y\,|\,x) + \|\nabla_y q(y\,|\,x)\| &= q(y\,|\,x)\left(1 + \|\nabla_y \log q(y\,|\,x)\|\right) \\ &= \exp\left(-a(x) - b(y) + c\langle x, y\rangle\right)\left(1 + \|b'(y) + x\|\right) \\ &\leqslant \exp\left(-a(x) + c\langle x, y\rangle\right) \exp\left(-b(y)\right)\left(1 + \|b'(y)\|\right)\left(1 + \|x\|\right). \end{aligned}$$

Consider $y \in \mathcal{Y}$. Since $\mathcal{Y}$ is open, there is r_0 such that for $0 < r < r_0$ the closed ball $\mathcal{V} = B(y, r)$ is a neighborhood of y satisfying $\mathcal{V} \subset \mathcal{Y}$. By the local boundedness of b', b is locally Lipschitz hence locally bounded hence there is r_1 such that $C(\mathcal{V}) := \sup_{y' \in \mathcal{V}} \exp\left(-b(y)\right)\left(1 + \|b'(y)\|\right) < \infty$ for

[9]For $u, v \in \mathcal{H}$, $v^T : \mathcal{H} \to \mathbb{R}$ denotes the linear form $x \mapsto \langle v, x\rangle$, and $uv^T : \mathcal{H} \to \mathcal{H}$ the linear operator $x \mapsto \langle v, x\rangle u$.

any $0 < r < r_1$. For $0 < r < \min(r_0, r_1)$ we have $\sup_{y' \in \mathcal{V}} \exp\left(c\langle x, y' - y\rangle\right) = \exp\left(cr\|x\|\right) < \infty$, hence for any $x \in \mathcal{X}$

$$\begin{aligned} C_{q,\mathcal{V}}(x) &:= \sup_{y' \in \mathcal{V}} \left\{ q(y' \,|\, x) + \|\nabla_y q(y' \,|\, x)\| \right\} \\ &\leqslant C(\mathcal{V}) \sup_{y' \in \mathcal{V}} \exp\left(-a(x) + c\langle x, y' - y\rangle + c\langle x, y\rangle\right) (1 + \|x\|) \\ &\leqslant C(\mathcal{V}) \exp\left(-a(x) + cr\|x\| + c\langle x, y\rangle\right) (1 + \|x\|) \,. \end{aligned}$$

By assumption (24) we have $\mathbb{E}_{X \sim P}\left\{(1 + \|X\|) C_{q,\mathcal{V}}(X)\right\} < \infty$ when $r > 0$ is small enough, i.e. (33) holds. Moreover, since $q(y \,|\, x) > 0$ for any $x \in \mathcal{X}$ we have $q_P(y) > 0$ i.e. $y \in \mathcal{Y}_P$, showing that $\mathcal{Y} = \mathcal{Y}_P$. By Lemma 8, f_P is differentiable at y with differential given by (34). Finally, by (23) we have

$$(x' - x)\left(\nabla_y^T \log q(y \,|\, x') - \nabla_y^T \log q(y \,|\, x)\right) = c(x' - x)(x' - x)^T \succeq 0, \quad \forall\, x, x' \in \mathcal{X}$$

hence $Df_P(y)$ is the expectation of a symmetric positive semi-definite operator. As a result, $Df_P(y)$ is symmetric positive semi-definite. As this holds for any $y \in \mathcal{Y}$, and since $\mathcal{Y}$ is open and convex, by Theorem 2, f_P implements a proximity operator, i.e., there exists a function φ_P such that $f_P(y) \in \operatorname{prox}_{\varphi_P}(y)$ for all $y \in \mathcal{Y}$.

Now, assume that $c > 0$ and consider $u \in \mathcal{H}$ such that $\langle u, Df_P(y)u\rangle = 0$. As

$$\langle u, Df_P(y)u\rangle = \tfrac{c}{q_P^2(y)} \mathbb{E}_{X \sim P} \mathbb{E}_{X' \sim P} q(y|X) q(y|X') \langle u, (X' - X)\rangle^2.$$

and $q(y|x)q(y|x')\langle u, x' - x\rangle^2 \geqslant 0$ for any x, x', by Markov's inequality we obtain that

$$q(y|X)q(y|X')\langle u, X' - X\rangle^2 = 0$$

almost surely on the draw of (X, X'). As $q(y|x) > 0$ for any x, this implies that $\langle u, X' - X\rangle = 0$ almost surely, hence there exists $d \in \mathbb{R}$ such that $\langle u, X\rangle = d$ almost surely. Since we assume that $P(\langle u, X\rangle = d) < 1$ for any nonzero $u \in \mathcal{H}$, it follows that $u = 0$. This shows that $Df_P(y) \succ 0$. We conclude using Corollary 2.

A.5. **Proof of Lemma 8.** By (32), $0 \leqslant q(y \,|\, x) \leqslant C(x)$ for any $y \in \mathcal{V}, x \in \mathcal{X}$, hence the numerator in (18), $n(y) := \mathbb{E}_{X \sim P}\left\{X q(y \,|\, X)\right\}$, is well-defined. As, $\|\nabla_y q(y \,|\, x)\|_{\mathcal{H}} \leqslant C(x)$, $\mathbb{E}_{X \sim P}\left\{X \nabla_y q(y \,|\, X)\right\} < \infty$ is similarly well-defined. Denote $\Delta(x, y, h) := q(y + h \,|\, x) - q(y \,|\, x) - \langle \nabla_y q(y \,|\, x), h\rangle$. As $|\Delta(x, y, h)| \leqslant 2C(x)\,\|h\|$ for any $x \in \mathcal{X}, y \in \mathcal{V}$ and h with $\|h\|_{\mathcal{H}}$ small enough, and as $\lim_{\|h\| \to 0} x\Delta(x, y, h) = 0$ for any $x \in \mathcal{X}, y \in \mathcal{V}$, by the dominated convergence theorem it follows that for any $y \in \mathcal{V}$

$$\lim_{\|h\|_{\mathcal{H}} \to 0} n(y + h) - n(y) - \mathbb{E}_{X \sim P}\left\{X \langle \nabla_y q(y \,|\, X), h\rangle\right\} = 0$$

showing that n is differentiable on $\mathcal{V}$ with differential $Dn(y) = \mathbb{E}_{X \sim P}\left\{X \nabla_y^T q(y \,|\, X)\right\}$. Similar arguments exploiting (32) show that q_P is differentiable on $\mathcal{V}$ with differential $Dq_P(y) = \nabla^T q_P(y)$ where $\nabla q_P(y) = \mathbb{E}_{X \sim P}\left\{\nabla_y q(y \,|\, X)\right\}$. In particular, q_P is continuous on $\mathcal{V}$, hence $\mathcal{V}_P = q_P^{-1}((0, \infty))$ is open.

For $y \in \mathcal{V}_P$, the denominator in (18) is $q_P(y) > 0$. By standard calculus f_P is differentiable at y and

$$\begin{aligned} Df_P(y) &= \frac{q_P(y)Dn(y) - n(y)Dq_P(y)}{q_P^2(y)} \\ &= \frac{\mathbb{E}_X\ \left\{X\nabla_y^T q(y\,|\,X)\right\} \cdot \mathbb{E}_{X'}\ \{q(y\,|\,X')\} - \mathbb{E}_X\ \{Xq(y\,|\,X)\} \cdot \mathbb{E}_{X'}\ \left\{\nabla_y^T q(y\,|\,X')\right\}}{q_P^2(y)} \\ q_P^2(y)Df_P(y) &= \mathbb{E}_{X,X'}\ \left\{X\left(q(y\,|\,X')\nabla_y^T q(y\,|\,X) - q(y\,|\,X)\nabla_y^T q(y\,|\,X')\right)\right\}. \end{aligned}$$

Now, we distinguish two cases:

- for x, x' such that $q(y\,|\,x)q(y\,|\,x') > 0$ using that $\nabla_y \log q_P = (\nabla_y q_P)/q_P$ where $q_P > 0$ we write

$$q(y\,|\,x')\nabla_y^T q(y\,|\,x) - q(y\,|\,x)\nabla_y^T q(y\,|\,x') = q(y\,|\,x)q(y\,|\,x')\left(\nabla_y^T \log q(y\,|\,x) - \nabla_y^T \log q(y\,|\,x')\right); \tag{35}$$

- for x, x' such that $q(y\,|\,x)q(y\,|\,x') = 0$, we have $q(y\,|\,x)$ or (non-exclusive) $q(y\,|\,x') = 0$. For example assume $q(y\,|\,x) = 0$. As $y' \mapsto q(y'\,|\,x)$ is non-negative, it is locally minimum at $y' = y$, and as it is differentiable this implies $\nabla_y q(y\,|\,x) = 0$. Similarly if $q(y\,|\,x') = 0$ then $\nabla_y q(y\,|\,x') = 0$. As a result (35) remains valid with the convention $\nabla_y \log q_P = 0$ where $q_P = 0$.

With the above observations we rewrite

$$\begin{aligned} q_P^2(y)Df_P(y) &= \mathbb{E}_{X,X'}\Big(Xq(y\,|\,X)q(y\,|\,X')\left(\nabla_y^T \log q(y\,|\,X) - \nabla_y^T \log q(y\,|\,X')\right)\Big) \\ &= \mathbb{E}_{X,X'}\Big(-Xq(y\,|\,X)q(y\,|\,X')\left(\nabla_y^T \log q(y\,|\,X') - \nabla_y^T \log q(y\,|\,X)\right)\Big) \qquad (36) \\ &\overset{X\leftrightarrow X'}{=} \mathbb{E}_{X,X'}\Big(-X'q(y\,|\,X')q(y\,|\,X)\left(\nabla_y^T \log q(y\,|\,X) - \nabla_y^T \log q(y\,|\,X')\right)\Big) \\ &= \mathbb{E}_{X,X'}\Big(X'q(y\,|\,X)q(y\,|\,X')\left(\nabla_y^T \log q(y\,|\,X') - \nabla_y^T \log q(y\,|\,X)\right)\Big) \qquad (37) \end{aligned}$$

We conclude by taking the half sum of (36) and (37).

A.6. **Proof of Proposition 4.** Since $p_{Y|X}$ is of the form (12) we have $a(x) + b(y) = F(x - y) + c\langle x, y\rangle$ for any $x, y \in \mathcal{H}$. A consequence is that $a(x) = F(x) - b(0)$ and $b(y) = F(-y) - a(0)$ for any x, y. In particular $a(0) + b(0) = F(0)$. It follows that $F(x - y) + c\langle x, y\rangle = a(x) + b(y) = F(x) + F(-y) - a(0) - b(0) = F(x) + F(-y) - F(0)$ for any x, y. Denoting $G(z) = G(z) - G(0)$, we get $G(x-y) + c\langle x, y\rangle = G(x) + G(-y)$. Specializing to $x = y$ yields $G(0) + c\|x\|^2 = G(x) + G(-x)$ for any x, hence $G(0) = 0$ and $c\|x\|^2 = G(x) + G(-x)$ for all x. Denoting $A(z), B(z)$ the odd and the even part of $G(z)$ we get $B(z) = c\|z\|^2/2$ and

$$A(x-y)+c\|x-y\|^2/2+c\langle x, y\rangle = G(x-y)+c\langle x, y\rangle = G(x)+G(-y) = A(x)+A(-y)+c\|x\|^2/2+c\|y\|^2/2$$

for any x, y. Thus, $A(x - y) = A(x) + A(-y)$ for any x, y and, as A is C^0 (by the continuity of F) it follows that A is linear: there is $\mu \in \mathcal{H}$ such that $A(x) = -c\langle x, \mu\rangle$ for all x, so that $G(z) = c\|z\|^2/2 - c\langle z, \mu\rangle = \frac{c}{2}\left(\|z - \mu\|^2 - \|\mu\|^2\right)$ and $F(z) = c\|z - \mu\|^2 + d$ with $d \in \mathbb{R}$. As the noise is centered, we conclude that $\mu = 0$.

A.7. **Proof of Lemma 2.** Equivalently we show that $G(x)-\log(1+|x|)$ is bounded from below.

Since $\int_{\mathbb{R}} e^{-G(x)}dx < \infty$ and G is convex, there is $a \in \mathbb{R}$ such that G is non-increasing on $(-\infty, a]$ and non-decreasing on $[a, +\infty)$ with $\lim_{|x|\to\infty} G(x) = +\infty$. By convexity again, it follows that there are $x_0 < a < x_1$ and $u_0 < 0 < u_1$ such that $G(x) \geqslant G(x_1) + u_1(x - x_1)$ and $G(x) \geqslant G(x_0) + u_0(x - x_0)$ for any $x \in \mathbb{R}$. On $[x_1, +\infty)$ we have $G(x) - \log(1 + |x|) \geqslant G(x_1) + u_1(x - x_1) - \log(1 + |x|)$ hence

$$\inf_{x\in[x_1,\infty)} \{G(x) - \log(1 + |x|)\} \geqslant G(x_1) - u_1x_1 + \inf_{x\geqslant x_1} \{u_1x_1 - \log(1 + |x|)\} > -\infty.$$

Similarly $\inf_{x\in(-\infty,x_0]}\{G(x) - \log(1 + |x|)\} > -\infty$. Finally, as G is convex on $[x_0, x_1]$ it is continuous on this compact interval hence $\inf_{x\in[x_0,x_1]}\{G(x) - \log(1 + |x|) = \min_{x\in[x_0,x_1]}\{G(x) - \log(1 + |x|)\} > -\infty$. Putting the pieces together establishes the result.

A.8. **Worked example: scalar denoising with Laplacian noise and Laplacian prior.** Consider $p_{Y|X}(y|x) \propto \exp(-|y - x|) =: q(y \,|\, x)$ and $p_X(x) = c\exp(-c|x|)/2$. Consider $y > 0$ (the case $y < 0$ is treated similarly by symmetry): we have

$$|y - x| + c|x| \quad = \quad \begin{cases} y - (1 + c)x & \text{if } x < 0 \\ y - (1 - c)x & \text{if } 0 \leqslant x \leqslant y \\ (1 + c)x - y & \text{if } x > y \end{cases}$$

hence for $c \neq 1$ we have

$$\begin{aligned}
\tfrac{2}{c}q_P(y) &= \int_{-\infty}^{+\infty} \exp(-|y - x| - |x|)\, dx \\
&= \int_{-\infty}^{0} e^{(1+c)x-y}dx + \int_{0}^{y} e^{(1-c)x-y}dx + \int_{y}^{+\infty} e^{y-(1+c)x}dx \\
&= e^{-y}\int_{-\infty}^{0} e^{(1+c)x}dx + e^{-y}\int_{0}^{y} e^{(1-c)x} + \int_{0}^{+\infty} e^{y-(1+c)(y+x)}dx \\
&= e^{-y}\int_{0}^{+\infty} e^{-(1+c)x}dx + e^{-y}\int_{0}^{y} e^{(1-c)x} + e^{-cy}\int_{0}^{+\infty} e^{-(1+c)x}dx \\
&= \tfrac{e^{-y}+e^{-cy}}{1+c} + e^{-y}\ \tfrac{e^{(1-c)y}-1}{1-c} = \tfrac{e^{-y}+e^{-cy}}{1+c} + \tfrac{e^{-cy}-e^{-y}}{1-c}
\end{aligned}$$

$$\begin{aligned}
\tfrac{2}{c}\int_{-\infty}^{+\infty} xq(y\,|\,x)p_X(x)dx &= \int_{-\infty}^{0} x\ e^{(1+c)x-y}dx + \int_{0}^{y} x\ e^{(1-c)x-y}dx + \int_{y}^{+\infty} x\ e^{y-(1+c)x}dx \\
&= -e^{-y}\ \int_{0}^{+\infty} x\ e^{-(1+c)x}dx + e^{-y}\ \int_{0}^{y} x\ e^{(1-c)x}dx \\
&\quad + \int_{0}^{+\infty} (y+x)\ e^{y-(1+c)(y+x)}dx \\
&= \left(e^{-cy}-e^{-y}\right)\ \int_{0}^{+\infty} x\ e^{-(1+c)x}dx + e^{-y}\ \int_{0}^{y} x\ e^{(1-c)x}dx \\
&\quad + ye^{-cy}\ \int_{0}^{+\infty} e^{-(1+c)x}dx \\
&= \tfrac{e^{-cy}-e^{-y}}{(1+c)^2} + \tfrac{ye^{-y}}{1+c} + e^{-y}\ \left[\left(x-\tfrac{1}{1-c}\right)\ \tfrac{e^{(1-c)x}}{1-c}\right]_0^y \\
&= \tfrac{e^{-cy}-e^{-y}}{(1+c)^2} + \tfrac{ye^{-y}}{1+c} + \tfrac{((1-c)y-1)\ e^{-cy}+e^{-y}}{(1-c)^2}
\end{aligned}$$

These derivations allow to express analytically $f(y) := \mathbb{E}(X|Y=y) = \frac{\int_{-\infty}^{+\infty} xq(y\,|\,x)p_X(x)dx}{q_P(y)}$.

References

[1] Arash Amini, Ulugbek Kamilov, Emrah Bostan, and Michael A Unser. Bayesian Estimation for Continuous-Time Sparse Stochastic Processes. *IEEE Trans. Signal Processing*, 61(4):907–920, 2013.

[2] Arash Amini, Michael A Unser, and Farokh Marvasti. Compressibility of Deterministic and Random Infinite Sequences. *IEEE Trans. Information Theory*, 59(11):5193–5201, 2011.

[3] Arindam Banerjee, Xin Guo 0001, and Hui Wang 0003. On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. Information Theory*, 51(7):2664–2669, 2005.

[4] Murat Belge, Misha Kilmer, and Eric Miller. Wavelet domain image restoration with adaptive edge-preserving regularization. *IEEE Trans. Image Process.*, 9(4):597–608, 2000.

[5] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[6] Martin Burger and Felix Lucka. Maximum a posteriori estimates in linear inverse problems with log-concave priors are proper Bayes estimators. *Inverse problems*, 30(11):114004–22, October 2014.

[7] Remi Gribonval. Should Penalized Least Squares Regression be Interpreted as Maximum A Posteriori Estimation? *IEEE Transactions on Signal Processing*, 59(5):2405–2410, 2011.

[8] Remi Gribonval, Volkan Cevher, and Michael E Davies. Compressible Distributions for High-Dimensional Statistics. *IEEE Trans. Information Theory*, 58(8):5016–5034, 2012.

[9] Remi Gribonval and Pierre Machart. Reconciling "priors" and "priors" without prejudice? In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 2193–2201, 2013.

[10] Rémi Gribonval and Mila Nikolova. A characterization of proximity operators. `https://hal.inria.fr/hal-01835101`, July 2018.

[11] T Helin and M Burger. Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems. *Inverse problems*, 31(8):085009, August 2015.

[12] S M Kay. *Fundamentals of Statistical Signal Processing˜: Estimation Theory*. Signal Processing. Prentice Hall, 1993.

[13] Abbas Kazerouni, Ulugbek Kamilov, Emrah Bostan, and Michael A Unser. Bayesian Denoising - From MAP to MMSE Using Consistent Cycle Spinning. *IEEE Signal Process. Lett.*, 20(3):249–252, 2013.
[14] Cécile Louchet and Lionel Moisan. Posterior Expectation of the Total Variation Model: Properties and Experiments. *SIAM J. Imaging Sci.*, 6(4):2640–2684, January 2013.
[15] P. Mathieu, M. Antonini, M. Barlaud, and I. Daubechies. Image coding using wavelet transform. *IEEE Trans. Image Process.*, 1(2):205–220, 1992.
[16] Jean-Jacques Moreau. Proximité et dualité dans un espace Hilbertien. *Bull. Soc. math. France*, 93:273–299, 1965.
[17] Mila Nikolova. Model distortions in Bayesian MAP reconstruction. *Inverse Problems and Imaging*, 1(2):399–422, 2007.
[18] Michael A Unser and Pouya D Tafti. *An introduction to sparse stochastic processes*. Cambridge University Press, 2014.